1-1-2016

Pirnas Variants And Lung Cancer Risk: A Post-Gwas Study

Rui Ye
Yale University, rui.ye@yale.edu

piRNAs variants and lung cancer risk: a post-GWAS study

Rui Ye

M.P.H. Thesis, Class 2016

Department of Environmental Health Sciences

Yale University School of Public Health

First Reader: Zhu Yong, PhD

Second Reader: Vasilis Vasiliou, PhD

ABSTRACT

Background: Lung cancer is the most frequently diagnosed cancer and one of the top cause of cancer death worldwide. The recent discovery of PIWI-interacting RNAs (piRNAs), one type of non-coding RNAs, has been shown to be involved in tumorgenesis pathways of various types of cancer types by accumulating evidences. However, the role of piRNAs in lung cancer development is underexplored.

Methods: Genotype and phenotype data were obtained from a genome-wide association study of lung cancer risk (3,702 cases and 3,739 controls) and 1,173 imputed piRNAs variants were analyzed by association analysis for lung cancer risk. A secondary expression analysis is also performed for 200 piRNAs to compare their expression levels in 497 lung adenocarcinoma patients versus 46 normal controls. Following *in vitro* functional analysis was also performed for the piRNAs variants identified from the association analysis.

Results: In the association analysis, one SNP (rs1169347) which can be mapped to two piRNAs (piR-5247 and piR-5671) was identified as statistically significantly associated with lung caner risk (FDR P-value <0.05). In the following functional analysis, results from cell viability assay showed a cell-growth-promoting effect of the major allele of the identified SNP. In the secondary expression analysis, 5 piRNAs were found significantly over-expressed in lung adenocarcinoma patients.

Conclusions: This comprehensive post-GWAS study provided important evidences for the role of piRNAs in lung cancer development through either their own functions or interactions with other protein-coding genes.

Introduction

Lung cancer is the most frequently diagnosed cancer and the first and second leading cause of cancer death among males and females worldwide[1]. In the United States, there will be an estimation of 224,390 new cases and 158,080 new deaths of lung cancer in 2016[2]. Moreover, the 5-year relative survival rate for lung cancer was only 18.4% from 2005 to 2011[3].

Currently, for non-small-cell lung cancer patients, which accounts for 85% of lung cancer cases, the main treatment options are surgery, radiotherapy and adjuvant chemotherapy[4]. However, since each treatment has its unavoidable side effects, a breakthrough in lung cancer treatment to increase the survival rate as well as improve the quality of life for lung cancer patients is needed. The advent of targeted therapy makes it possible to reduce the toxicity to patients compared to cytotoxic drugs[5]. Therefore, to discover novel agents with clinical significance that can be served as target for treatment of lung cancer in the future is especially important.

In recent years, increasing evidences suggested that the non-protein-coding portion of the genome is of crucial functional importance for disease development, including cancer[6]. Many studies suggested that the non-coding RNAs (ncRNAs) function through modulation of transcriptional or posttranscriptional processes[7]. Such transcriptional and posttranscriptional modifications would lead to a highly conserved pathway in which the small non-coding RNAs (sncRNAs) bind to protein complexes (PPD or Argonaute) and form the RNA-induced silencing complexes (RISC) to inhibit the expression of its target sequences[8]. The main small silencing RNAs can be classified into 3 categries: small interfering RNAs (siRNAs), microRNAs (miRNAs) and PIWI-interacting RNAs (piRNAs)[9].

The length of piRNA sequence is between 26 and 31 nucleotides (nt), slightly longer than siRNAs and miRNAs (between 21 and 26 nt)[10]. The primary function of piRNAs is to stabilize the germ line genome by silencing the transposon elements (TEs) through a highly conserved pathway which doesn't require Dicer during the process, while miRNAs or siRNAs-induced silencing pathways require Dicer[9]. Besides the TE-silencing function of piRNAs in the germ line, a growing number of studies are investigating its role in somatic cells. The detection of 4 main types of Piwi proteins (PIWIL1/HIWI, PIWIL2/HILI, PIWIL 3, and PIWIL4) in mammalian somatic tissues provides evidence for the existence of somatic piRNAs[11]. There are two pathways for the biogenesis of somatic piRNAs. In the primary processing pathway, long piRNA precursors are transcribed from piRNA clusters, cleaved and midified in the cytoplasma, and then transported into the nucleus loaded with Aubergine (AUB) or PIWI proteins[11]. In the amplification loop (ping-pong cycle), which is activated by piRNA-induced silencing complexes (piRISCs) produced in the primary pathway, piRNAs are modified and amplified to target on active TEs through a slicer-mediated cleavage[9]. Moreover, the PIWI-piRNAs pathway can regulate the transposon loci or even non-transposon loci outside the germline tissues through histone modifications and DNA methylation[12].

Recent studies found that piRNAs and piRNA-like transcripts are involved in tumorigenesis in a range of tumor types[13]. Both oncogenic and tumor-suppressing roles of piRNAs have been found by microarray screening, next generation sequencing (NGS), and real-time quantitative reverse transcription-polymerase (PCR) chain reaction analyses[14]. Several preliminary studies discovered the over-expression of PIWI proteins in several tumor types, such as seminomas, breast cancer, cervical cancer, glioma, colon cancer, etc[15,16,17,18,19]. One possible mechanism proposed by Siddiqi et.al[20] was that the presence of piRNAs and PIWI proteins in the

cancer tissues would result in aberrant DNA methylation and over-silencing of the promoting regions of tumor suppressor genes, and then trigger the tumorgenesis. Therefore, piRNAs have high potential to be a new prognostic biomarker or new therapy target for various types of cancer.

To our knowledge, there are very few previous studies investigating the association between piRNAs and lung cancer risk. Therefore, the aim of this study was to examine whether piRNAs variants are associated with lung cancer risk. We used the data of 3,817 cases and 3,921 controls from three cohort studies[21]. In addition, we further tested the role of the identified single nucleotide polymorphisms (SNPs) through *in vitro* functional analysis. Finally, we also identified several piRNAs that are significantly different expressed in lung adenocarcinoma compared to normal lung tissues by conducting a secondary analysis using the data derived from a scientific report of piRNAs expression profiling in several tumor types[22].

Materials and Methods

Study population and data

The population of this study is derived from a genome-wide association study of lung cancer[21], in which the subjects are from three cohort studies – Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study (ATBC)[23], the Prostate, Lung, Colon, Ovary Screening Trial (PLCO)[24], and the Cancer Prevention Study II Nutrition Cohort (CPS-II)[25]. The accessible individual genotype and phenotype data are downloaded from Database of Genotypes and phenotypes (dbGaP, Study Accession: phs000336.v1.p1) to a secure server at Yale University and decrypted and extracted according to dbGaP guidelines. The total population of this study is 7738, with 3,817 cases and 3,921 controls.

In the secondary expression data analysis, 252 piRNAs for 497 lung adenocarcinoma patients and 46 controls are obtained from the supplemental table 2 of the scientific report[22]. The unit of piRNAs expression is defined as reads per kilobase per million mapped reads (RPKM) and all RPKM values are obtained from the scientific report.

Data Cleaning and Management

All processes of data cleaning and management are performed by PLINK v 1.07[26]. Subjects from the three studies are genotyped on one of the four platforms (Illumina Human240K300K, HumanHap550K, Human610Quadv1, and Human1M-Duov3). The data of 539,000 SNPs in 3,817 cases and 3,921 controls are merged into one complete dataset. 124 pairs of subjects were found to have a familial relationship by identity by by descent (IBD) analysis (pi-hat ($\hat{\pi}$) >= 0.2) and each member of the 124 pairs was excluded from the final analysis. The dataset was restricted to SNPs with call rate ≥ 90% and Hardy-Weinberg Equilibrium test (HWE) $P > 0.0001$. Then, principal components analysis (PCA) was carried out using EIGENSTRAT[27] and 173 subjects were excluded as outliers. Finally, the sample file contained 533,002 SNPs in 3702 cases and 3739 controls.

For the expression data, 52 piRNAs were excluded from the final analysis because 13 of these piRNAs are mapped with overlaps to microRNAs (miRNAs), 36 are mapped to small nucleolar RNAs (snoRNAs), and 3 are mapped to transfer RNAs (tRNAs). After confirming the correct mapping by piRNABank and UCSC genome browser, 200 piRNAs were included in the final analysis.

piRNA Variant Genotype Imputation

The piRNA SNP list including the copy number and genome loci is obtained from piRNABank[28]. SNPs with copy number >100 were excluded because evidence showed that piRNAs with lower copy number are more likely to be involved in the regulation of protein-coding gene expression[29]. The 1,000 Genomes Phase 3 haplotype data were used as reference panel for imputation[30]. The imputation was performed using IMPUTE v2.3.1 software[31]. Fine mapping was performed through imputation of all SNPs with minor allele frequency (MAF) >1% in 5MB segments and all the coordinates information are collected from Genome Reference Consortium GRCh37/hg19 on USCS genome browser[32].

Association Analyses

The statistical analysis of the association study was performed by SNPTEST v2.5[33] using unconditional regression regression and the additive allele model, controlling for sex, age, original study participation, genotyping platform, and the first two principal components. The number of the principal components to control is determined by the study-wide genomic inflation-factor (GIF) and the corresponding QQ plot. The odds ratio (OR), 95% confidence interval (95% CI), nominal p-value, and false discovery rate-adjusted (FDR) P-value is provided for every association. The Manhattan plot and QQ plot are generated by R using qqman package[34].

Comparison of piRNAs expression level between normal and lung adenocarcinoma samples

A scatter plot visualizing the different expression level of the 200 included piRNAs between samples from 497 lung adenocarcinoma patients and 46 controls was created by J-

Express software[35]. The 2-tail t-test was used to detect the difference of individual piRNA expression level between samples from 497 lung adenocarcinoma patients and 46 controls for the 200 included piRNAs. Bonferroni-adjustment for multiple comparisons has been applied.

Materials for Functional Analyses

Human lung carcinoma epithelial cells (A549) are purchased from ATCC and are maintained in F-12K medium supplemented with 10% fetal bovine serum (FBS) and 1% penicillin. The wild type piRNA mimics of the two identified piRNAs, piRNA-5247 and piRNA-5671, were purchased from IDT, and the negative control non-targeting RNA were purchased from QIAGEN. Cells were transfected using LipofectAMINE RNAiMAX transfection reagent. The transfection rate was close to 100% as detected in our lab, and the toxicity of the transfection reagent showed little impact on the cell growth.

Cell Viability Assay

Cells are treated as triplicates for piRNA-5247, piRNA-5671 and the negative control RNA in the 96-well plate and the cell viability was evaluated by the CellTiter 96 AQueous One Solution Cell Proliferation Assay (MTS) kit (Promega). In one transfection experiment setting, 3 readings will be recorded by a microplate spectrophotometer at an absorbance of 490nm at 48, 72 and 96 hours, respectively. The different impact of the three RNAs on cell viability was determined by 2-tail t-test.

Results

Two identified piRNAs associated with lung cancer risk

The baseline characteristics of the included 3,702 cases and 3,739 controls are showed in Table 1. There are more males in control group than case group. The age distribution is similar between the two groups. A larger proportion of controls are from PLCO study while more cases are from ATBC study. And samples from the cases are mostly genotyped on HumanHap550K and Human610Quadv1 array whereas controls are genotyped on all the four arrays.

After all the data cleaning processes, genotype data of 533,002 SNPs from a total population of 7,441 have been included in the PCA analysis. A total of 1,173 SNPs that can be mapped to piRNAs of our interest are successfully imputed and included into the final association studies. The association between these 1,173 variants and lung cancer risk is displayed in a Manhattan Plot (Figure 1a). After adjusting for multiple comparisons by Bonferroni-correction, only one SNP (rs11639347) is statistically significant associated with lung cancer risk. rs11639347 can be mapped to two overlapping piRNAs, piR-5247 and piR-5671. As showed in Table 2, the minor allele of rs11639347 is a risky allele that increases lung cancer risk with an odds ratio (OR) of 1.17 (95% confidence interval (CI): 1.09, 1.27). Information about the SNP name, mapped piRNAs, position, allele, minor allele frequency, OR, nominal P-value, and FDR P-value for the top 3 identified SNPs are also included in Table 2. The association analysis is controlled for sex, age, original study participation, genotyping platform, and the first two principal components. A QQ plot (Figure 1b) demonstrating the control for covariates included in the association analysis is provided. Moreover, a plot (Figure 2) displaying the PCA results for the 2 principal components in which original studies are annotated in different colors is also provided.

Functional role of the major allele of the two identified piRNAs

Wild-type mimics for piR-5247 and piR-5671 were delivered into A549 cell lines. A non-targeting RNA mimics were also delivered into A549 cell lines as negative controls. Figure 3 showed the sequence information of piR-5247 and piR-5671. The allele locus of rs11639347 was highlighted as well. The transfection of each mimics has been done in triplicates. Figure 4a shows the reading results at 48, 72, and 96 hours after transfection of the 3 mimics. Interestingly, even if the wild-type allele of the two identified piRNAs has a protective role as indicated by the association analysis, the cells transfected with piR-5247 and piR-5671 can still significantly promote the A549 cell growth compared to negative-control RNA-treated A549 cells. Similarly, the cell growth trend showed in Figure 4b also revealed a same function of the wild-type allele.

Individual piRNAs Expression Level Difference

The scatter plot (figure 5) shows the mean expression level of each individual piRNAs among 497 lung adenocarcinoma patients and 46 normal controls. Most piRNAs have very low expression level in both lung adenocarcinoma and normal samples. However, the expression level was detectable in several outlying piRNAs which showed different expression patterns in tumor samples compared to normal samples. The seven highest expressed piRNAs in tumor samples have been listed in table 3. The information about the piRNA name, position, coding region, mean expression level in normal samples, mean expression level in lung tumor samples, nominal P-values generated by 2-tail t-test, and FDRP-values are provided. From table 2, 5 piRNAs (piR-14620, piR-2732, piR-51809, piR-19521, and piR-15232) are statistically significantly different expressed between normal and tumor samples. Among them, piR-14620 is of the highest expression level and all of the 5 piRNAs are up-regulated in tumor samples. The

only piRNA that is down-regulated in tumor samples of the top 7 piRNAs was piR-31637. However, after Bonferroni correction, the difference of its expression level was not statistically significant.

Discussion

This is the first comprehensive post-GWAS study combining the results of association analysis, the functional analysis and the expression profiling analysis to explore the involvement of piRNAs in lung cancer development. From the association analysis, we have identified the variant in one SNP (rs11639347) that is significantly associated with the increase risk of lung cancer. The location of the variant (Chromosome 15: 79024350) and the 2 piRNAs, piR-5247 (Chromosome 15: 79024333-79024361) and piR-5671 (Chromosome 15: 79024327-79024355) is in intergenic region. This suggests that the functional changes caused by the 2 piRNAs may be attributed to their own functions. In addition, we established a follow-up *in vitro* functional analysis to further explore the role of the two piRNAs in A549 cell growth. The cell viability assay result shows a cell-growth promoting effect of the protective allele of rs11639347 on A549 lung cancer cell lines, suggesting that this SNP may play an oncogenic role in lung cancer development. Moreover, we used the same experiment setting to transfect the U87 cells with the 2 piRNAs mimics, a glioma cell line. The result (data not shown) showed that the 2 piRNAs have no effect on the growth of U87 cells. Therefore, these data suggest that the SNP rs11639347 may be tissue-specific involved in the lung cancer development pathway. Further functional analysis, such as cell viability assay, colony formation assay which is regarded as the "gold standard" cellular-sensitivity assay[36], and expression profiling should be conducted for the risk allele (rs11639347).

From the expression analysis, we have identified 5 piRNAs that are up-regulated in lung adenocarcinoma samples. Among which, piR-14620, the highest expressed piRNA, is located in the intron of gene KIAA0825. piR-2732 is located in the intron of gene RPL3, which encodes the ribosome proteins and is involved in DNA repair through regulation of p21 function[37]. piR-51809 is located in the intron of gene CPA6. piR-19521 is located in the intergenic region. piR-15232 is located in the exon of HIST1H2BJ, which encodes H2B histone protein[38]. Therefore, future studies are needed to explore the role of KIAA0825, CPA6 and HIST1H2BJ in lung cancer development. Functional analysis of piR-2732 should be further conducted since it seems be involved in cancer development through regulation of DNA repair and cell apoptosis.

There are several strengths of this study. First, in association study, the use of 1,000 Genome Phase 3 haplotype reference panel guarantees a wide coverage of piRNA embedded SNPs during imputation. Second, the result of cell viability assay shows rs11639347 only functions to promote the lung cancer cell growth, suggesting it may be a good specific target for future lung cancer treatment. Third, from the association study and expression analysis, we've identified several piRNAs variants associated with lung cancer risk are located in protein-coding regions as well as intergenic regions. This finding provides further evidences to suggest that piRNAs play important roles in tumorgenesis through either their independent biological roles or interactions with oncogenes or tumor-suppressor genes. Lastly, the combination of association study, expression analysis, and functional analysis provides a comprehensive understanding of the identified SNPs that are associated with lung cancer risk.

While this study suggests a role of piRNAs in lung cancer development, there are some limitations. First, even if the reference panel has a very wide coverage of SNPs, piRNAs variants that are not included in the panel cannot be imputed. Therefore, a small part of piRNAs may be

missing in this study. Second, in the expression analysis, we've found that some piRNAs included in the original scientific report are actually overlapped with other types of small non-coding RNAs, leaving an inevitable potential mapping quality issue in our secondary analysis of the expression data. However, we believe this issue has been minimized since we checked the genomic loci of all the individual piRNAs and excluded those overlapped with other types non-coding RNAs. Third, due to the limited baseline characteristics of study participants provided by dbGaP datasets, we are only able to control for the confounding effects of sex, age, original study participation, genotyping platform.

Future work will focus on the functional analysis for the identified piRNAs variants from both association and expression analysis. More specifically, for rs11639347, we will continue analyze the function role of its risk allele in lung cancer cell-growth by cell-viability assay and clonogenic assay. In addition, we will perform expression profiling of the 2 piRNAs embedded in this SNP. Finally, we will further investigate the mechanism of the pathway it is involved in lung tumorgenesis on protein-interaction level.

In conclusion, this comprehensive post-GWAS study provides strong evidence for an association between piRNAs and lung cancer risk. The *in vitro* functional analysis provides further evidences of an oncogenic role of piRNAs in lung cancer development. Future functional work will help us to better understand the function of the identified piRNAs that may be involved in lung tumorgenesis.

Acknowledgement

Foremost, I would like to express my gratitude to my thesis first-reader professor Yong Zhu. His guidance helped me in all the time of research and writing of this thesis with his patience, knowledge and encouragement. Secondly, I would like to acknowledge professor Vasilis Vasiliou as my thesis second-reader. His office is always open and I am gratefully indebted to him for his very valuable comments on this thesis. Besides my thesis advisors, I would also like to thank Daniel Jacobs, who is a Ph.D. student at professor Zhu's lab, for his continuous supports and inputs in all aspects of this work.

Tables

Table 1: A table of baseline characteristics of included association study population. The number of participants of each category for and the percentage for sex, age, study and array is provided included. The percentage of each category in cases and controls is also included.

Characteristics		Cases (N=3702)	Controls (N=3739)
Sex			
	Male (%)	2873 (77.61%)	3279 (87.70%)
	Female (%)	829 (22.39%)	460 (12.30%)
Age			
	<50 (%)	6 (0.16%)	6 (0.16%)
	50-54 (%)	457 (12.34%)	427 (11.42%)
	55-59 (%)	923 (24.93%)	865 (23.13%)
	60-64 (%)	1130 (30.52%)	1112 (29.74%)
	65-69 (%)	853 (23.04%)	909 (24.31%)
	70-74 (%)	310 (8.37%)	386 (10.32%)
	75+ (%)	23 (0.62%)	34 (0.91%)
Study			
	PLCO	1311 (35.41%)	1819 (48.65%)
	CPSII	663 (17.91%)	650 (17.38%)
	ATBC	1728 (46.68%)	1270 (33.97%)
Array			
	Human240K300K	0 (0%)	965 (25.81%)
	HumanHap550K	757 (20.45%)	829 (22.17%)
	Human610Quadv1	2945 (79.55%)	1797 (48.06%)
	1M-Duov3	0 (0%)	148 (3.86%)

Table 2: A table for summary of top-3 identified piRNAs embedded SNPs that are associated with lung cancer risk.

SNP	piRNAs[1]	Position	Minor/Common Allele	MAF (cases/controls)	OR[2]	95% CI	Nominal P-value	FDR P-value[3]
rs11639347	piR-5247 piR-5671	Chr15: 79024350	T/C	0.41/0.38	1.17	(1.09, 1.27)	3.560E-05	0.042
rs13382748	piR-21626	Chr2: 95450931	C/T	0.11/0.10	1.26	(1.12,1.43)	2.190E-04	0.257
rs60534722	piR-16828	Chr12: 24554473	A/G	0.17/9.19	0.85	(0.77,0.94)	1.498E-03	1.757

[1]: Identified SNPs are located within the genome loci of the piRNAs; [2]: Odds ratio for the minor allele associated with lung cancer; [3]: Bonferroni-correction for 1173 comparisons.

Table 3: A table for summary data of top-7 identified piRNAs from the expression analysis.

piRNAs[1]	Name	Position	Strand	Gene[2]	Mean-Normal[3]	Mean-Tumor[4]	Nominal P-Value	FDR P-Value[5]
FR043670	piR-14620	Chr5: 93905174-93905200	-	Intron of KIAA0825	486.94	1025.32	6.280E-05	0.001
FR090905	piR-20009	Chr7: 145694484-145694511	+	Intergenic	389.33	711.72	0.047	9.391
FR082269	piR-31637	ChrM: 619-650	+	Intergenic	358.28	149.19	0.005	1.090
FR205579	piR-2732	Chr22: 39709883-39709914	-	Intron of RPL3	26.71	140.45	1.060E-18	2.120E-16
FR038165	piR-51809	Chr8:68497704-68497734	-	Intron of CPA6	3.22	59.26	2.300E-16	4.610E-14
FR111727	piR-19521	Chr11:10530940-10530967	-	Intergenic	6.07	48.22	6.540E-23	1.310E-20
FR197889	piR-15232	Chr6:27100537-27100567	+	Exon of HIST1H2BJ	4.56	40.98	5.720E-41	1.140E-38

[1]: piRNAs name used in the scientific report; [2]: The genome region where piRNAs are located; [3,4]: The mean expression level (RPKM) of tumor and control samples. [5]: Bonferroni-correction for 200 comparisons.

Figures

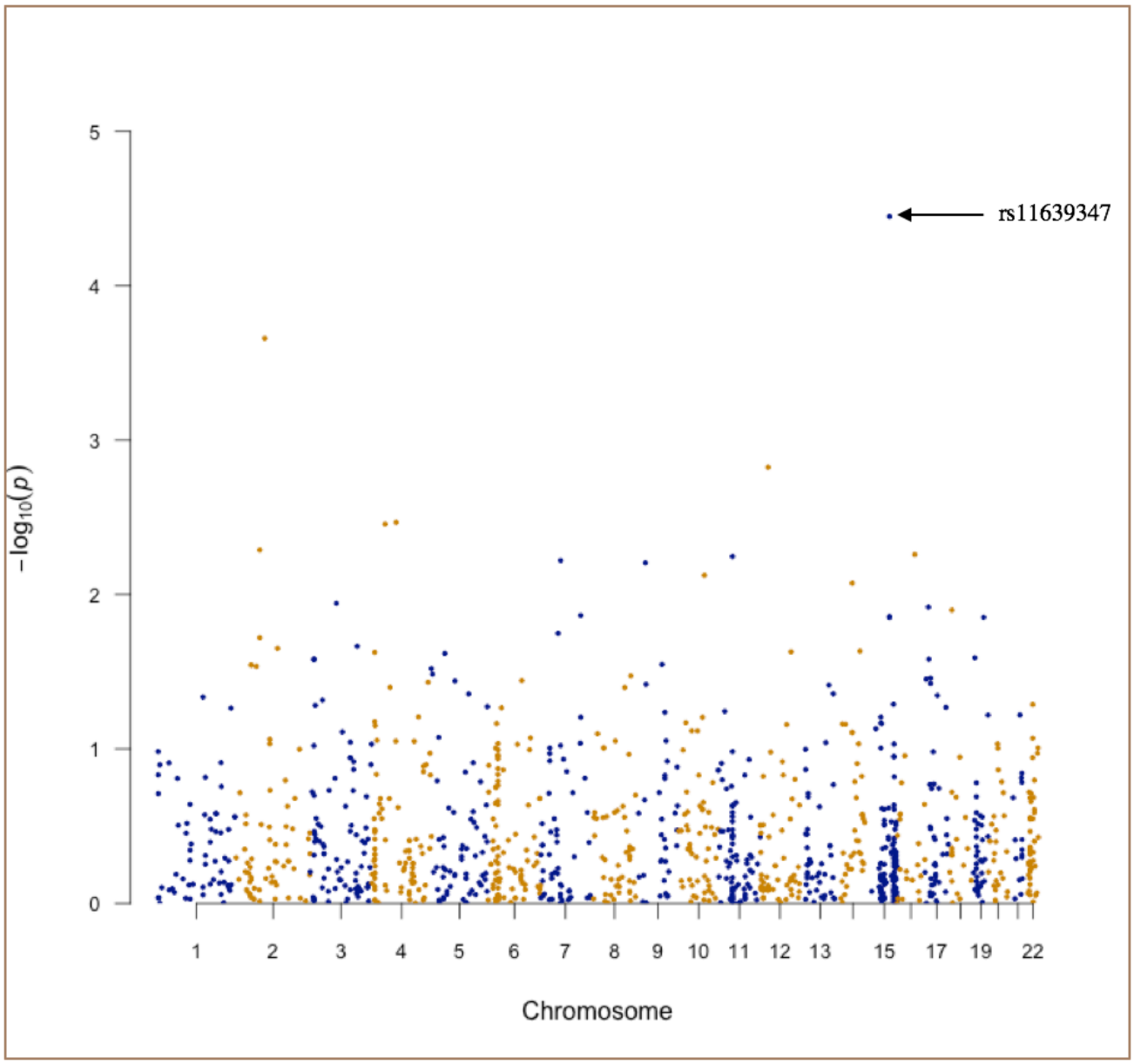

Figure 1a: A Manhattan plot displaying the results from association study for the 1,173 piRNAs variants. The variant rs1169347 is annotated in the plot.

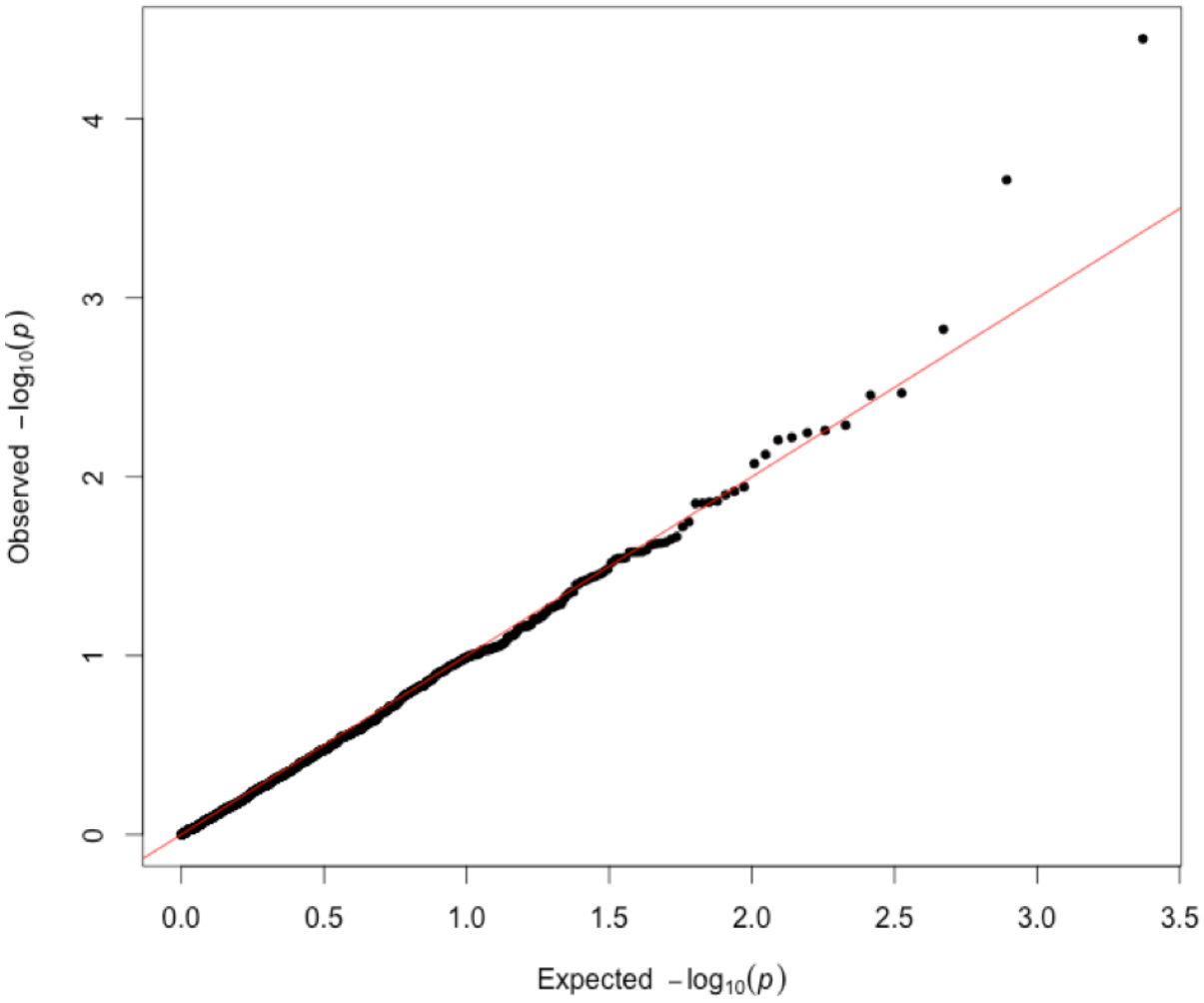

Figure 1b: A QQ plot indicating the adjustment for covariates from the association study.

Figure 2: A plot showing the first two principal components categorized by original studies from the principal components (PCA) analysis.

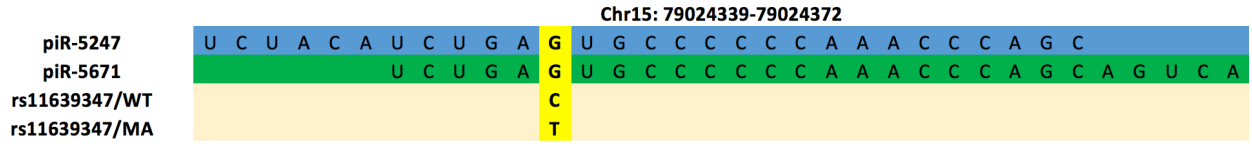

Figure 3: Sequence and genomic locus information for the wild-type and minor allele of rs11639347 as well as the 2 embedded piRNAs (piR-5247 and piR5671).

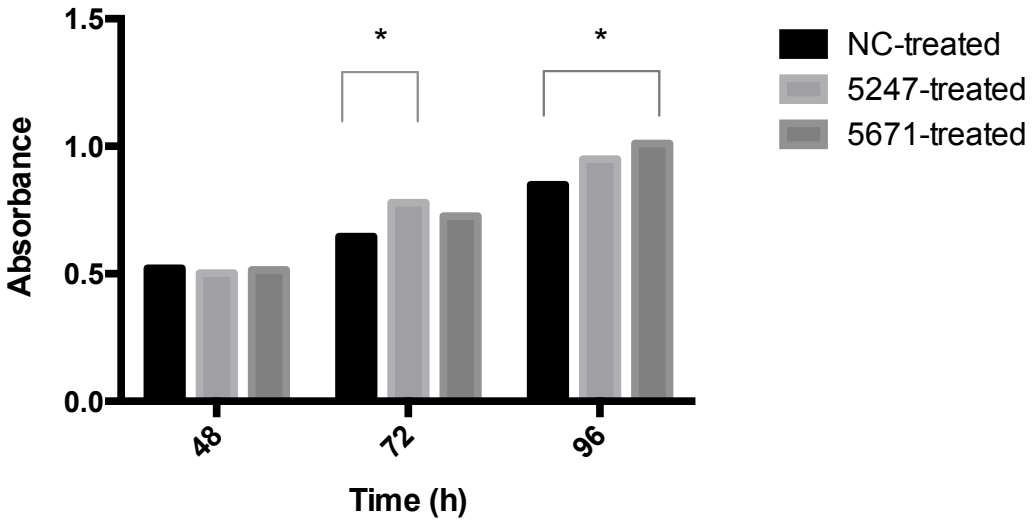

Figure 4a: The results of cell viability assay readings for NC-treated, piR-5247-treated and piR-5671-trated A519 cells at 48, 72, and 96 hours respectively.

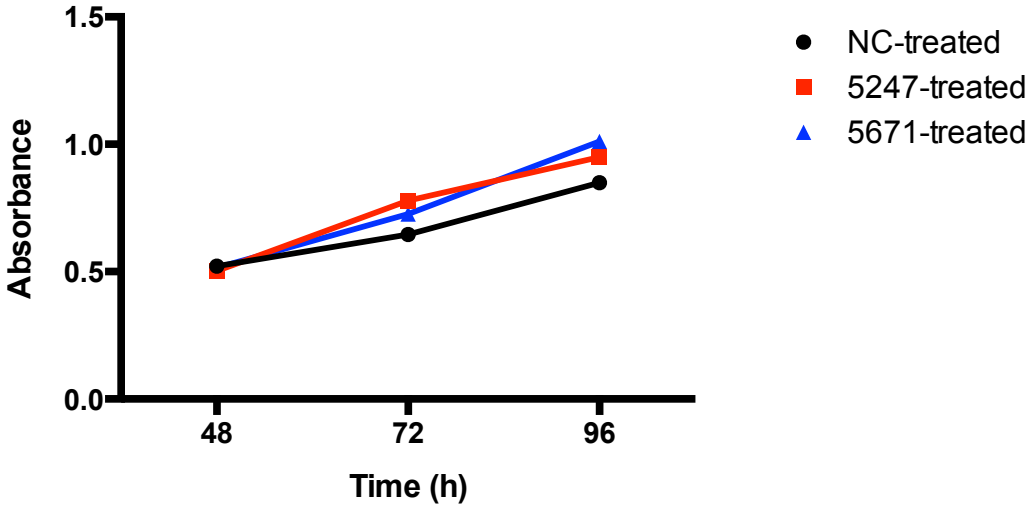

Figure 4b: The time trend of A549 cell growth showing the effect of piR-5247 and piR-5671.

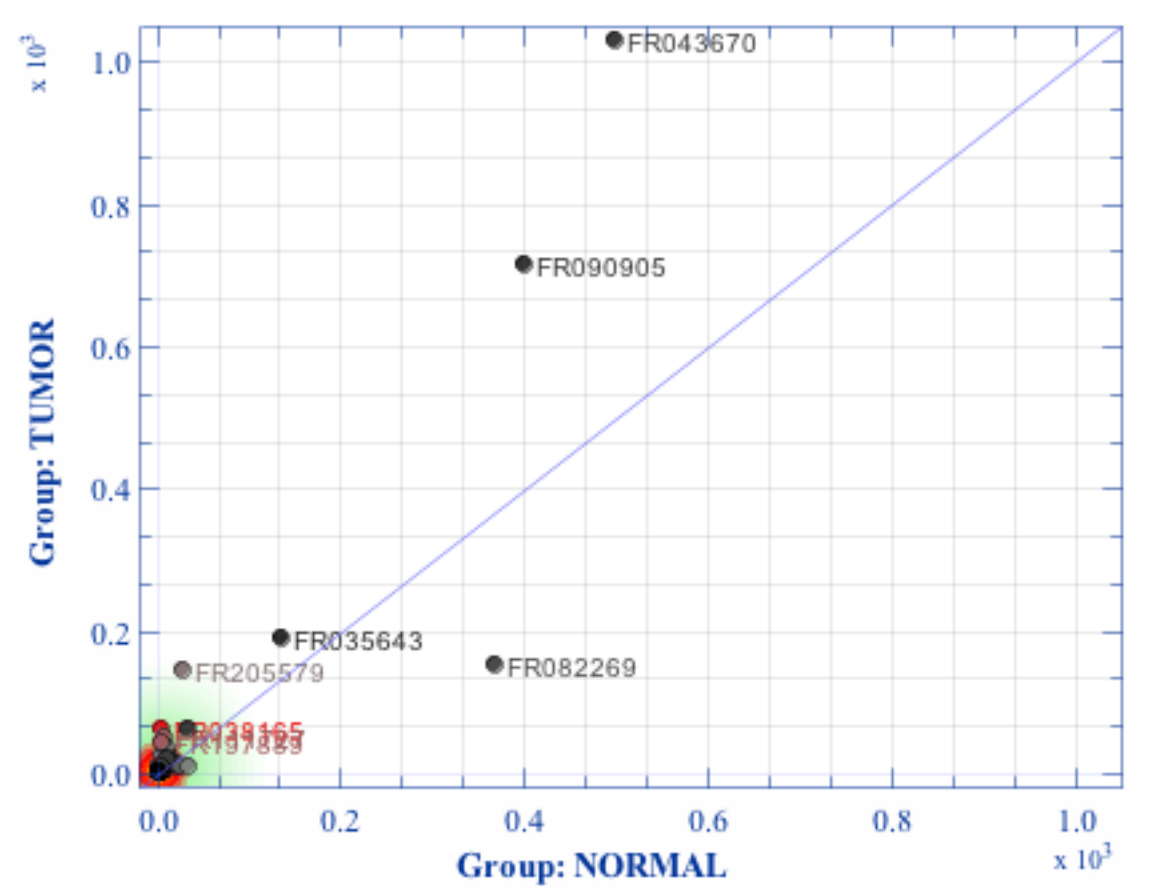

Figure 5: A Scatter Plot displaying the results of the secondary expression. X-axis is the mean expression level of each piRNAs in 46 normal samples while Y-axis is the mean expression level of each piRNAs in 497 tumor samples

Reference

[1] Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., & Jemal, A. (2015). Global cancer statistics, 2012. *CA: a cancer journal for clinicians*, *65*(2), 87-108.

[2] Society, A. C. (2016). Cancer facts and figures 2016.

[3] Howlader N, Noone AM, Krapcho M, Garshell J, Miller D, Altekruse SF, Kosary CL, Yu M, Ruhl J, Tatalovich Z,Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). SEER Cancer Statistics Review, 1975-2012, National Cancer Institute.

[4] Pallis, A. G. (2011). A Review of Treatment in Non-small-cell Lung Cancer. *Eur. Respir. Dis*, *7*(1), 27-31

[5] Ricciardi, Serena, Silverio Tomao, and Filippo de Marinis. "Toxicity of targeted therapy in non–small-cell lung cancer management." *Clinical lung cancer* 10.1 (2009): 28-35.

[6] Mercer, Tim R., Marcel E. Dinger, and John S. Mattick. "Long non-coding RNAs: insights into functions." *Nature Reviews Genetics* 10.3 (2009): 155-159.

[7] Lee, T. I., & Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. *Cell*, *152*(6), 1237-1251.

[8] Almeida, R., & Allshire, R. C. (2005). RNA silencing and genome regulation.*Trends in cell biology*, *15*(5), 251-258.

[9] Siomi, M. C., Sato, K., Pezic, D., & Aravin, A. A. (2011). PIWI-interacting small RNAs: the vanguard of genome defence. *Nature reviews Molecular cell biology*,*12*(4), 246-258.

[10] Zaratiegui, M., Irvine, D. V., & Martienssen, R. A. (2007). Noncoding RNAs and gene silencing. *Cell*, *128*(4), 763-776.

[11] Ross, R. J., Weiner, M. M., & Lin, H. (2014). PIWI proteins and PIWI-interacting RNAs in the soma. *Nature*, *505*(7483), 353-359.

[12] Peng, J. C., & Lin, H. (2013). Beyond transposons: the epigenetic and somatic functions of the Piwi-piRNA mechanism. *Current opinion in cell biology*, *25*(2), 190-194.

[13] Esteller, Manel. "Non-coding RNAs in human disease." *Nature Reviews Genetics* 12.12 (2011): 861-874.

[14] Mei, Yuping, David Clark, and Li Mao. "Novel dimensions of piRNAs in cancer."*Cancer letters* 336.1 (2013): 46-52.

[15] Lee, J. H., Jung, C., Javadian-Elyaderani, P., Schweyer, S., Schütte, D., Shoukier, M., ... & Mantilla, A. (2010). Pathways of proliferation and antiapoptosis driven in breast cancer stem cells by stem cell protein piwil2.*Cancer research*, *70*(11), 4569-4579.

[16] Liu, J. J., Shen, R., Chen, L., Ye, Y., He, G., Hua, K., ... & Barsky, S. H. (2010). Piwil2 is expressed in various stages of breast cancers and has the potential to be used as a novel biomarker. *Int J Clin Exp Pathol*, *3*(4), 328-337.

[17] Wang, Y., Liu, Y., Shen, X., Zhang, X., Chen, X., Yang, C., & Gao, H. (2012). The PIWI protein acts as a predictive marker for human gastric cancer. *Int J Clin Exp Pathol*, *5*(4), 315-325.

[18] Lee, J. H., Schütte, D., Wulf, G., Füzesi, L., Radzun, H. J., Schweyer, S., ... & Nayernia, K. (2006). Stem-cell protein Piwil2 is widely expressed in tumors and inhibits apoptosis through activation of Stat3/Bcl-XL pathway. *Human molecular genetics*, *15*(2), 201-211.

[19] Sun, G., Wang, Y., Sun, L., Luo, H., Liu, N., Fu, Z., & You, Y. (2011). Clinical significance of Hiwi gene expression in gliomas. *Brain research*, *1373*, 183-188.

[20] Siddiqi, S., & Matushansky, I. (2012). Piwis and piwi-interacting RNAs in the epigenetics of cancer. *Journal of cellular biochemistry*, *113*(2), 373-380.

[21] Landi, M. T., Chatterjee, N., Yu, K., Goldin, L. R., Goldstein, A. M., Rotunno, M., ... & Bergen, A. W. (2009). A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *The American Journal of Human Genetics*, *85*(5), 679-691.

[22] Martinez, V. D., Vucic, E. A., Thu, K. L., Hubaux, R., Enfield, K. S., Pikor, L. A., ... & Lam, W. L. (2015). Unique somatic and malignant expression patterns implicate PIWI-interacting RNAs in cancer-type specific biology.*Scientific reports*, *5*.

[23] ATBC Cancer Prevention Study Group. (1994). The alpha-tocopherol, beta-carotene lung cancer prevention study: design, methods, participant characteristics, and compliance. *Annals of epidemiology*, *4*(1), 1-10.

[24] Hayes, R. B., Sigurdson, A., Moore, L., Peters, U., Huang, W. Y., Pinsky, P., ... & Hoover, R. N. (2005). Methods for etiologic and early marker investigations in the PLCO trial. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *592*(1), 147-154.

[25] Calle, E. E., Rodriguez, C., Jacobs, E. J., Almon, M. L., Chao, A., McCullough, M. L., ... & Thun, M. J. (2002). The American cancer society cancer prevention study II nutrition cohort. *Cancer*, *94*(9), 2490-2501.

[26] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), 559-575.

[27] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, *38*(8), 904-909.

[28] Lakshmi, S. S., & Agrawal, S. (2008). piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic acids research*,*36*(suppl 1), D173-D177.

[29] Fu, A., Jacobs, D. I., & Zhu, Y. (2014). Epigenome-wide analysis of piRNAs in gene-specific DNA methylation. *RNA biology*, *11*(10), 1301-1312.

[30] 1000 Genomes Project Consortium. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56-65.

[31] B. Howie, J. Marchini, and M. Stephens (2011) Genotype imputation with thousands of genomes. *G3: Genes, Genomics, Genetics* 1(6): 457-470.

[32] Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC.*Genome research*, *12*(6), 996-1006.

[33] J. Marchini and B. Howie (2010) Genotype imputation for genome-wide association studies. Nature Reviews Genetics.

[34] Turner, S.D. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. biorXiv DOI: 10.1101/005165 (2014).

[35] Dysvik, B., & Jonassen, I. (2001). J-Express: exploring gene expression data using Java. *Bioinformatics*, *17*(4), 369-370.

[36] Langdon, S. P. (2004). *Cancer cell culture: methods and protocols* (Vol. 88). Springer Science & Business Media.

[37] Esposito, D., Crescenzi, E., Sagar, V., Loreni, F., Russo, A., & Russo, G. (2014). Human rpL3 plays a crucial role in cell response to nucleolar stress induced by 5-FU and L-OHP. *Oncotarget*, *5*(22), 11737.

[38] Zhong, R., Roeder, R. G., & Heintz, N. (1983). The primary structure and expression of four cloned human histone genes. *Nucleic acids research*,*11*(21), 7409-7425.